




## SOFTWARE TOOL ARTICLE

**REVISED** CASAS: Cancer Survival Analysis Suite, a web based application [version 2; referees: 2 approved]Manali Rupji <sup>1</sup>, Xinyan Zhang<sup>1</sup>, Jeanne Kowalski<sup>1,2</sup><sup>1</sup>Winship Cancer Institute of Emory University, Atlanta, GA, 30322, USA<sup>2</sup>Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, 30322, USA**v2** First published: 15 Jun 2017, 6:919 (doi: [10.12688/f1000research.11830.1](https://doi.org/10.12688/f1000research.11830.1))  
Latest published: 31 Jul 2017, 6:919 (doi: [10.12688/f1000research.11830.2](https://doi.org/10.12688/f1000research.11830.2))**Abstract**

We present CASAS, a shiny R based tool for interactive survival analysis and visualization of results. The tool provides a web-based one stop shop to perform the following types of survival analysis: quantile, landmark and competing risks, in addition to standard survival analysis. The interface makes it easy to perform such survival analyses and obtain results using the interactive Kaplan-Meier and cumulative incidence plots. Univariate analysis can be performed on one or several user specified variable(s) simultaneously, the results of which are displayed in a single table that includes log rank p-values and hazard ratios along with their significance. For several quantile survival analyses from multiple cancer types, a single summary grid is constructed. The CASAS package has been implemented in R and is available via <http://shinygispa.winship.emory.edu/CASAS/>. The developmental repository is available at <https://github.com/manalirupji/CASAS/>.

This article is included in the **RPackage** gateway.**Open Peer Review**Referee Status:  

Invited Referees		
	1	2
<b>REVISED</b>		
<b>version 2</b>	report	report
published 31 Jul 2017		
<b>version 1</b>		
published 15 Jun 2017	report	

- 1 **Gang Han**, Texas A&M University School of Public Health, USA
- 2 **Seon-Young Kim** , Korea Research Institute of Bioscience and Biotechnology, Korea, South

**Discuss this article**

Comments (0)

**Corresponding author:** Jeanne Kowalski ([jeanne.kowalski@emory.edu](mailto:jeanne.kowalski@emory.edu))

**Author roles:** **Rupji M:** Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Zhang X:** Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Kowalski J:** Conceptualization, Resources, Software, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Rupji M, Zhang X and Kowalski J. **CASAS: Cancer Survival Analysis Suite, a web based application [version 2; referees: 2 approved]** *F1000Research* 2017, 6:919 (doi: [10.12688/f1000research.11830.2](https://doi.org/10.12688/f1000research.11830.2))

**Copyright:** © 2017 Rupji M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Grant information:** Research reported in this publication was supported in part by the Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI under award number P30CA138292.

**First published:** 15 Jun 2017, 6:919 (doi: [10.12688/f1000research.11830.1](https://doi.org/10.12688/f1000research.11830.1))

**REVISED Amendments from Version 1**

Figure 1, Figure 3 and Figure 4 have been updated to illustrate standard survival analysis, quantile survival analysis, tests for proportional hazards, and selection of a cut point for a continuous variable in a single example.

See referee reports

## Introduction

Kaplan-Meier (KM) estimates and the Cox Proportional Hazards model have gained huge popularity among clinicians when depicting survival trends and in identifying prognostic biomarkers in cancer research. There is a range of commercial software (SAS, STATA, SPSS, PRISM) available for researchers to carry out survival analysis. However, these programs have several disadvantages; commercial software is proprietary and involves restricted usage with rigid outputs, which cannot be changed easily. Open source software such as R is gaining popularity, but the user needs to learn programming skills, which may be very time consuming for clinicians and biomedical researchers with limited programming exposure.

Standard survival analysis involves a single cause of failure. However, in other cases, clinicians may encounter many other causes of failure in addition to a specific cause of interest. In such cases, a competing risk analysis needs to be carried out, where an individual is exposed to two or more causes of failure but its eventual failure is due to only one cause. While several packages are available to conduct competing risk survival analysis in R, making the right choice presents another layer of confusion to the user.

As opposed to traditional KM or Cox regression analysis, typically a risk factor measured at baseline is examined for its association with survival thereafter. During follow-up, however, things may have changes, such that either the effect of a fixed baseline risk factor may vary over time, resulting in a weakening or strengthening of associations over time or the risk factor itself may vary over time. In the former case, such as effect is often seen in what appears to be significant differences in survival, not necessarily overall and among all survival times, but early on or at later survival times. We address such time-dependent effects on survival by creating two additional tools, one for landmark<sup>1</sup> and another for quantile survival analysis<sup>2,3</sup>. As an example, the user may want to study the effect of chemotherapy on a specific cancer population and thus divides the data into a responder vs a non-responder group. The issue with this approach is that the responder cannot be deemed one, unless they survive until the time of response. In addition, being in the responder group gives them an unfair survival advantage leading to an immortal time bias. To overcome these issues, the investigator may need to perform a landmark analysis by removing the patients with an event (or censored) before the landmark time from the analysis.

Most tools are available as separate packages. As an alternative, CASAS provides a comprehensive survival analysis suite of tools commonly encountered in cancer research. By providing a GUI interface, the user can readily perform any number of these analyses

by simply uploading their data and selecting the variables relevant to the analysis.

In summary, CASAS suite of tools is a one-stop shop for conducting some of the most common survival analyses in cancer research without requiring any prior programming knowledge. It is a web-based application that, as a single tool, can carry out KM plot, univariate hazard ratio, landmark analysis, quantile survival analysis and competing risk analysis. It allows a user to combine results from various studies or cancer types as well.

## Methods

### KM survival analysis

Standard survival analysis uses the TCGA BRCA data for 553 patients that underwent radiotherapy<sup>4</sup> and uses the 'survminer' package to display the KM plots. Either categorical or continuous variables can be used for stratification. Continuous variables can be dichotomized by either the 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup> percentile or an optimal cut point. A log-rank test is used to estimate the overall differences between the survival curves (Figure 1). A single overall survival curve without stratification can be plotted using the 'All patients' option. The user can also test for validity of the proportional hazards assumption (a non-significant Schoenfeld residual test p-value based on the cox.zph function in the 'survival' package) in the univariate survival analysis tab.

### Competing risk survival analysis

Competing risk survival analysis is based on Fine and Gray's Model<sup>5</sup>, using the 'cmprsk' package in R. To illustrate, 35 AML/ALL patient's data who underwent Hematopoietic stem cell transplantation (HSCT) affected by either AML or ALL<sup>6</sup> are used. Cumulative incidence plot was plotted using CumIncidence.R function available in the package. This tool also displays the Gray's p-value based on the competing risk code (Figure 2). Either categorical or continuous group variables/'All patients' can be used.

### Landmark analysis

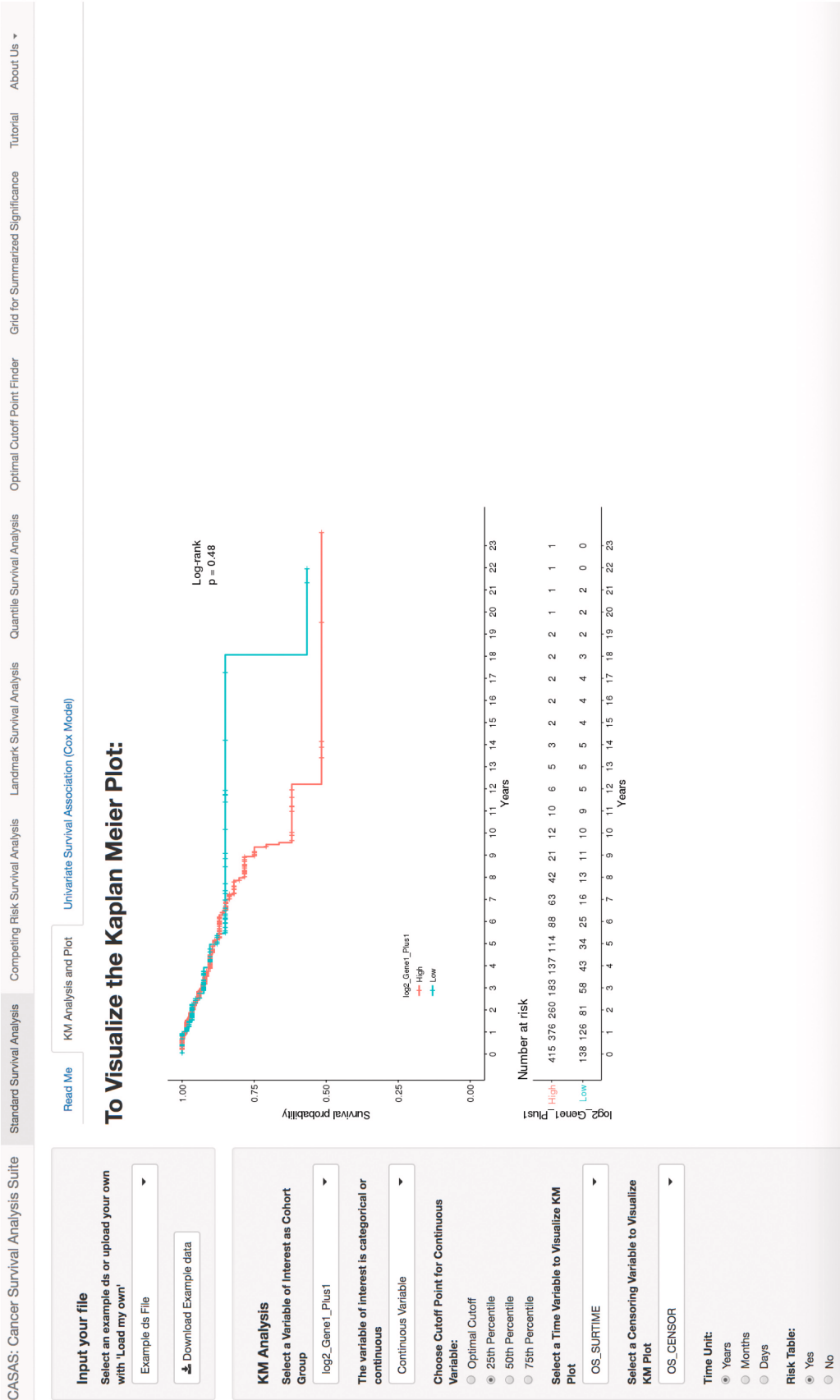
Landmark analysis is based on a user input landmark time. Stanford Heart transplant dataset is used<sup>7</sup>. The tool generates an overall KM plot and a landmark KM plot with log rank test p-values (Figure 3). The user can also opt for a CI curve instead of a KM plot and allows similar categorical/continuous variable inputs.

### Quantile survival analysis

Quantile survival analysis is based on method developed by 2,3, implemented in the 'cequre' package in R. The example data used is the 553 patient Expression with Clinical data for TCGA – BRCA patients given radiotherapy<sup>4</sup>. Survival time difference between the dichotomized continuous or categorical variable will be estimated with 95% CI for each quantile (Figure 4). A forest plot to represent quantile wise differences between the means and the overall differences is also provided as output.

## Implementation

The CASAS software is written in R and tested using version 3.3.0. The Interactive KM, CIF plots and data tables are made visible through a web browser using the shiny R package ([www.rstudio.com/shiny](http://www.rstudio.com/shiny)).



CASAS: Cancer Survival Analysis Suite

Standard Survival Analysis

Competing Risk Survival Analysis

Landmark Survival Analysis

Quantile Survival Analysis

Optimal Cutoff Point Finder

Grid for Summarized Significance

Tutorial

About Us

Read Me

CIF Analysis and Plot

Univariate Survival Association (Fine and Gray Model)

Input your file

Select an example ds or upload your own with 'Load my own!'

Example ds File

Download Example data

CIF Analysis

Select a Variable of Interest as Cohort Group

dis

The variable of interest is categorical or continuous

Categorical Variable

Select a Time Variable to Visualize CIF Plot

time

Select a Censoring Variable to Visualize CIF Plot

status

Input Time Points:

☐ Yes

☒ No

Time Points Input

0, 10, 20, 30

Time Unit:

☒ Years

☐ Months

☐ Days

Event Codes:

☒ 1

☐ 2

☐ 0

Censor Codes:

☒ 1

☐ 2

☐ 0

To Visualize the Cumulative Incidence Function Plot:

AML 1

AML 2

Gray's P-value: 0.079

Show 100 entries

Time (Months)

CIF Estimate (95% CI)

ALL 1

0 0.0598 (0.0035, 0.2421)

10 0.1179 (0.0182, 0.3186)

20 0.1765 (0.0399, 0.3929)

30 0.1765 (0.0399, 0.3929)

40 0.1765 (0.0399, 0.3929)

50 0.1765 (0.0399, 0.3929)

60 0.1765 (0.0399, 0.3929)

70 0.1765 (0.0399, 0.3929)

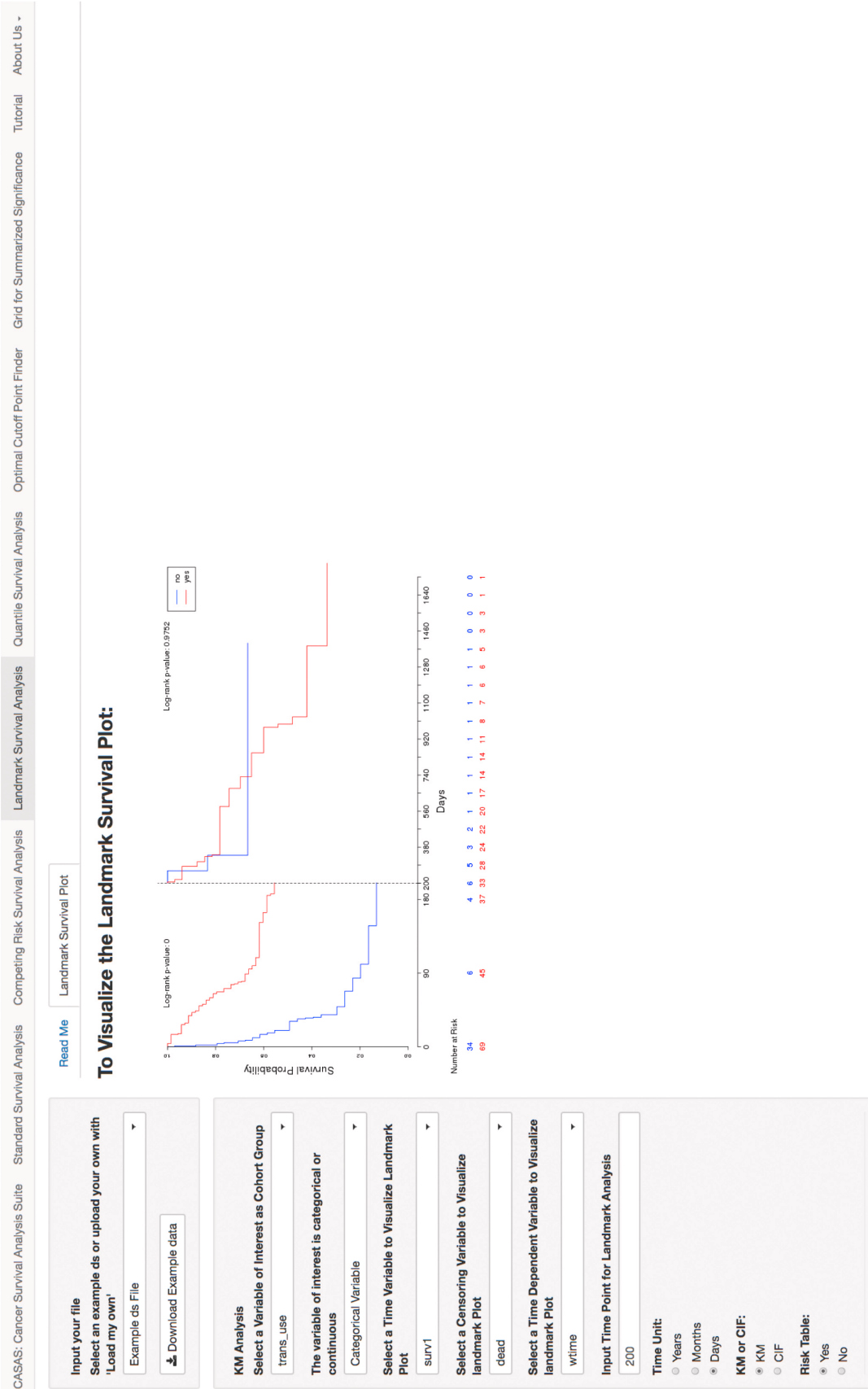
AML 1

0 0 (NaN, NaN)

10 0.3862 (0.1408, 0.601)

Figure 2. Competing risk analysis.

This tab includes competing risk analysis using the BMT data (<http://www.stat.unipg.it/luca/R/>). The left panel includes options to select variables for univariate survival association analysis based on Fine and Gray's model, and includes options to select variables for cumulative incidence function analysis. For the Univariate Analysis, the user chooses variable of interest to enter into the analysis as well as the event and the censor code. To generate a Cumulative Incidence Function (CIF) plot and table, the user can choose either a categorical variable to compare or 'All Patients' (Like in Figure 1). Users can also select the appropriate time unit based on the data and the time points of interest, where applicable. Censor code can also be specified. The univariate result table or plot is displayed on the right panel.





**Figure 4. Quantile survival analysis.** The quantile regression tab shows three plots based on the same data used in Figure 1. The first is the overall KM plot with number at risk table for overall survival and log rank test p-value. The second is the survival time difference with 95% CI between two dichotomized groups at the 10 quantiles Q1 to Q10 (defined as 10 percentile to 100 percentile by 10 percentile at mean survival time among all patients). The third plot is the summary using forest plot for the survival time difference at 10 quantiles. The overall in forest plot corresponds to the transformed HR and 95% CI for overall survival ( $\log [1/HR]$ ).

## Operation

Using a windows 7 Enterprise SP1 PC with a 32.0 GB RAM and an 3.30 GHz Intel® Xeon® Processor E5 Family, for 553 TCGA BRCA patients who received radiotherapy<sup>4</sup>, it took 2.26 s to create an interactive KM plot and 6.52 s to generate quantile analysis plots. Stanford heart transplant data<sup>7</sup> for landmark survival analysis generated the KM plots in about 1.10s. For data of 35 patients who underwent Hematopoietic stem cell transplantation (HSCT) affected by either AML or ALL<sup>6</sup>, it took 2.66s to display the CIF plot. The developmental repository is available at <https://github.com/manalirupji/CASAS/>.

Archived source code as at the time of publication is available at: <http://doi.org/10.5281/zenodo.832845><sup>8</sup>.

## Discussion

CASAS is a suite of tools that allows a user to conduct various types of survival analysis through an interactive application in R. We show by example, various types of cancer survival analyses that can be performed based on the questions of interest. Our tool will serve as a platform for many physicians and researchers to conduct preliminary analyses before heading to statisticians to conduct advanced analyses.

## Data and software availability

The CASAS web tool (<http://shinygispa.winship.emory.edu/CASAS/>) includes preprocessed example data under each tab. The user could use the example data or could upload a dataset of their choice in the same format as the example data. The tool will work best with data input as a .txt or .csv file. Time variable can be input in days, months or years and the appropriate time unit selection can be made. This tool accepts censor variables in various formats, for example, character annotation such as Dead/Alive, or integer variable 0/1, etc. The tool may throw errors if there are missing data within the censor variables but it can handle missing data left blank for the cohort variables. Use of any other characters for missing data may cause the program to throw errors.

For landmark analysis, data from 103 patients on the waitlist for Stanford Heart transplant program as available in the survival package is used<sup>7</sup>. The data can be accessed in R using `data(jasa)`. For both Kaplan Meier survival analysis and quantile survival analysis, Level 3 RNASeqV2 Breast Cancer (BRCA) data was downloaded from the TCGA data portal<sup>4</sup>. 553 patients that received radiotherapy and had survival information were used. Gene expression data for a specific biomarker gene was log2 transformed. The user has the choice to divide the data based on either the twenty-fifth percentile (set as default) or the fiftieth or the seventy-fifth, or an optimal cut point based on the martingale residuals. Similarly, for competing risk analysis, the user could choose the example data or upload their own. The example data consists of 35 patients with acute leukemia who underwent Hematopoietic stem cell transplantation (HSCT) affected by either AML or ALL<sup>6</sup> (<http://www.stat.unipg.it/luca/R>).

The developmental repository is available at <https://github.com/manalirupji/CASAS/>.

Archived source code as at the time of publication: <http://doi.org/10.5281/zenodo.832845><sup>8</sup>

License: CASAS is available under the GNU public license (GPL-3).

## Competing interests

No competing interests were disclosed.

## Grant information

Research reported in this publication was supported in part by the Biostatistics and Bioinformatics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI under award number P30CA138292. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

## References

1. Anderson JR, Cain KC, Gelber RD: **Analysis of survival by tumor response.** *J Clin Oncol.* 1983; **1**(11): 710–9.  
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Huang Y: **Restoration of monotonicity respecting in dynamic regression.** *J Am Stat Assoc.* 2017; **112**(518): 613–622. (In press).  
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Huang Y: **Quantile Calculus and Censored Regression.** *Ann Stat.* 2010; **38**(3): 1607–1637.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. NCI and NHGRI: **The Cancer Genome Atlas (TCGA) Data Portal.** Accessed in December 2015.  
[Reference Source](#)
5. Fine JP, Gray RJ: **A Proportional Hazards Model for the Subdistribution of a Competing Risk.** *J Am Stat Assoc.* 1999; **94**(446): 496–509.  
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Scrucca L, Santucci A, Aversa F: **Competing risk analysis using R: an easy guide for clinicians.** *Bone Marrow Transplant.* 2007; **40**(4): 381–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Crowley J, Hu M: **Covariance analysis of heart transplant survival data.** *J Am Stat Assoc.* 1977; **72**(357): 27–36.  
[PubMed Abstract](#) | [Publisher Full Text](#)
8. manalirupji: **manalirupji/CASAS: CASASv1.1.0.** *Zenodo.* 2017.  
[Data Source](#)



# Open Peer Review

Current Referee Status:



Version 2

Referee Report 23 August 2017

doi:10.5256/f1000research.13225.r25109



**Seon-Young Kim** 

Genome Structure Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea, South

The manuscript entitled 'CASAS: Cancer Survival Analysis Suite, a web based application' by Rupji *et al* describes a web server, a shiny R based tool for interactive survival analysis and visualization. They provide several kinds of survival analysis with detailed explanations and example data sets. The web application is easy to use and is likely to be useful for many researchers. I tested each of the examples and found that they work as described in the manuscript. If I add one comment, I found that the CASAS doesn't support analysis of multiple genes at once, which is necessary when one wants to identify prognostic gene(s) from omics data. The authors may discuss this point and refer to an example of web servers that support the analysis of multiple genes (i.e <http://appex.kr>).

## References

1. Kim SK, Hwan Kim J, Yun SJ, Kim WJ, Kim SY: APPEX: analysis platform for the identification of prognostic gene expression signatures in cancer. *Bioinformatics*. 2014; **30** (22): 3284-6 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 23 Aug 2017

**Manali Rupji**, Emory University, USA

We thank the referee for their kind review. The CASAS tool supports several univariate, single gene analyses, the results of which are able to be summarized altogether in a table. The tool offers a comprehensive analyses with respect to each gene under study. Thus, the CASAS tool implements a forward selection algorithm for identifying statistically significant individual gene associations with clinical outcome.

**Competing Interests:** No competing interests were disclosed.

Referee Report 01 August 2017

doi:10.5256/f1000research.13225.r24659



**Gang Han**

Department of Epidemiology and Biostatistics, Texas A&M University School of Public Health, College Station, TX, USA

The authors have answered all my comments. I have no more additional comments.

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 01 Aug 2017

**Manali Rupji**, Emory University, USA

We thank the referee for taking the time to review our work, and thank them for the kind review.

**Competing Interests:** No competing interests were disclosed.

---

Version 1

Referee Report 06 July 2017

doi:10.5256/f1000research.12784.r24053



**Gang Han**

Department of Epidemiology and Biostatistics, Texas A&M University School of Public Health, College Station, TX, USA

Authors of this article attempted to facilitate the routine survival analysis for clinicians and general health science professionals by building a web based survival analysis software with underlying R programming. Because survival data has increased usage in clinical research, this work is useful to fill out the gap of survival analysis for researchers without training in statistics and programming. But given a number of limitations, its practical value is questionable. Due to the following multiple missing components, this article does not look complete. This reviewer suggests major revision to improve its usability:

1. The first missing component is the test of model assumptions. The validity of the proportional hazard assumption was not described.
2. The second missing piece is the detailed description of required data format. For example, some practitioners may label event/censor as "1/0", while others will write "not censored/censored." What is the requirement for the data file header? For missing values, some may leave it blank, others may write "missing" or "unknown." These things are all trivial for statisticians but useful for clinical users to know and bear in mind. The authors may add a separate section to emphasize the correct format.
3. The third missing piece is a comprehensive example. Although some figures were shown, potential users may look for a detailed example to follow. The authors had made the point that this software can perform certain survival analyses, but they failed to explain/illustrate how to perform these analyses. A major revision is necessary to answer the question of "how" in an example that looks similar to majority of the users' data and analytical needs.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

No

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

No

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 20 Jul 2017

**Manali Rupji**, Emory University, USA

We thank the reviewer for their thoughtful comments that have led to an improved paper. Below, we respond individually to each comment.

*1. The first missing component is the test of model assumptions. The validity of the proportional hazard assumption was not described.*

We appreciate this reviewer's careful review of the tool and in bringing this point to our attention. A test for proportional hazards (PH) was included as part of the standard survival analysis option. We recognize that several methods exist for checking the proportional hazards assumption and results may vary depending on the method applied. In response to the reviewer's comment, we have included under the documentation tab, "Read Me", that the proportional hazards assumption is examined by applying the method of Schoenfeld residuals. If this method indicates a violation of the proportional hazards assumption as indicated by a significant test of PH (versus alternative of non-PH) based on such residuals, then other methods should be examined that are outside the scope of this tool since results may vary depending on the method implemented (Hiller *et al.* 2015).

Ref:

Hiller L, Marshall A, Dunn J. Assessing violations of the proportional hazards assumption in cox regression: does the chosen method matter? *Trials*. 2015; 16(suppl 2): P134.

*2. The second missing piece is the detailed description of required data format. For example, some practitioners may label event/censor as "1/0", while others will write "not censored/censored." What is the requirement for the data file header? For missing values, some may leave it blank, others may write "missing" or "unknown." These things are all trivial for statisticians but useful for clinical users to know and bear in mind. The authors may add a separate section to emphasize the correct format*

We appreciate the reviewer's point of clarity here and have addressed their recommendation of including a separate paragraph that highlights the required format. This new paragraph is listed under the, "Data and software availability" section in the paper.

*3. The third missing piece is a comprehensive example. Although some figures were shown, potential users may look for a detailed example to follow. The authors had made the point that this software can perform certain survival analyses, but they failed to explain/illustrate how to perform these analyses. A major revision is necessary to answer the question of "how" in an example that looks similar to majority of the users' data and analytical needs.*

In response to the reviewer's comment, we have updated the one example initially used for

illustrating the quantile survival analysis to also illustrate the following: standard survival analysis, a test of proportional hazards assumption, and selection of a cutpoint for a continuous variable. Additionally, we have updated the specific dataset for landmark analysis, to include the time corresponding to the time-dependent cohort variable. We have left the other previous examples provided in the initial submission to supplement the 'how to' of individual methods. While we recognize the importance of having a single example to streamline the methods of survival analysis, we also thought it important to emphasize that the methods are not dependent on each other and that they can be implemented as several stand-alone methods under one tool.

**Competing Interests:** No competing interests were disclosed.

---